A preliminary study on estimating word imageability labels using Web image data mining

Marc A. Kastner Ichiro Ide Yasutomo Kawanishi
Takatsugu Hirayama Daisuke Deguchi Hiroshi Murase
Nagoya University
kastnerm@murase.is.i.nagoya-u.ac.jp
{ide, kawanishi, murase}@i.nagoya-u.ac.jp
{takatsugu.hirayama, ddeguchi}@nagoya-u.jp

1 Introduction

Through many applications, Natural Language Processing (NLP) became ubiquitous in daily life; whether it is machine translation, personal assistants, search engines, or recommendation engines, NLP is a key element in many multimedia applications. However, connecting NLP to the human is an ongoing problem for various real-world applications. Despite progress in state-of-the-art methods, language processing can often feel disconnected from the user, a human, leading to *unnatural* results. This problem is commonly called *semantic gap*, often connected to cross-modality multimedia processing, and usually involves human perception. To understand the underlying semantics of NLP and multi-modal applications in real-world applications, the human perception of language needs to be taken in consideration.

Imageability is an idea from Psycholinguistics to quantize the human perception of words. On a scale from, in laymans terms, abstract to concrete, the ability to conceptualize a term as a mental image, is described with a number. Further research showed, that this relationship of language and imageability has further implications for language acquisition, language understanding, and the use of grammar. Therefore, it seems natural to put this research in an NLP context, and use it for multi-modal applications. However, many datasets used in this field are created in labor-intensive experiments, ranging from annotation by hand by test subjects in academic studies, to crowd-sourcing using Amazon Mechanical Turk[1].

In this research, we propose a method using image-based data-mining to make an estimation on the imageability of words. The key idea is the assumption that there is an intrinsic relationship between the imageability of words, how we perceive the world around us and, how we capture this in images we upload to social media platforms. Therefore, first, we crawl large image sets for words for which we have imageability ground truth labeling. Next, a data-mining approach using a set of low-level visual features is applied to all images. For each word, the visual features are used to train a model to predict imageability. The model is evaluated using a testing data set. The proposed method can be used to increase the corpus of imageability dictionaries.

In Section 2, previous research on imageability is discussed. Section 3 discusses our proposed method of using image data-mining for estimating the imageability of words. Obtaining the ground-truth data for the imageability annotations, as well as the images for the data-mining is discussed in Section 4. Lastly, Section 5 shows the preliminary results of our proposed method, before concluding the paper in Section 6.

2 Related work

The idea of imageability and human perception in language understanding goes back to the 1960s, starting in Psychology. From there, the concept naturally found its way into Psycholinguistics, NLP, and Computer Science.

Paivio et al.[5] first proposed the concepts of imageability, concreteness, and meaningfulness as measurements for human perception of natural language. Since then, there has been ongoing research, connecting language understanding and language acquisition to the imageability of words and concepts. For example, the imageability of verbs has implications on grammar usage for different contexts[7], which could provide helpful knowledge to create more natural language depending on context. There are imageability dictionaries for English [1][6] as well as other languages. However, the dictionary creation process is labor-intensive, as the annotations are commonly obtained through crowd-sourcing or user studies involv-

[1] https://www.mturk.com/

ing test subjects.

While the opportunities of imageability are not yet deeply researched for multimedia research, it finds its way into recent multi-modal approaches [9]. Further applications would be deeper understanding of semantics, improvement of image retrieval, and a greater understanding of black-box machine learning models (so-called Explainable AI.)

In previous research [3], we introduced a method to estimate the visual variety of terms with a data-driven approach to create ideal datasets, before doing a simplistic clustering-based approach on the visual features to determine the variety. While the direct connection of visual variety to imageability was left open for future research, a relationship between the ideas is undeniable. The evaluation covered a small number of 25 terms related to vehicles, which led to promising results. However, larger scale experiments turned out to be unfeasible due to constraints in the data acquisition process. The approach came with multiple downsides, most prominently: First, the proposed method being tied to the WordNet [4] hierarchy, as it relied on the homonym-hypernym relationship of words. Second, it used ImageNet [2] as an image source, which is rather narrow, and also limited to nouns. To the best of our knowledge, there has been no other research in this field using images to estimate imageability, visual variety, or related concepts.

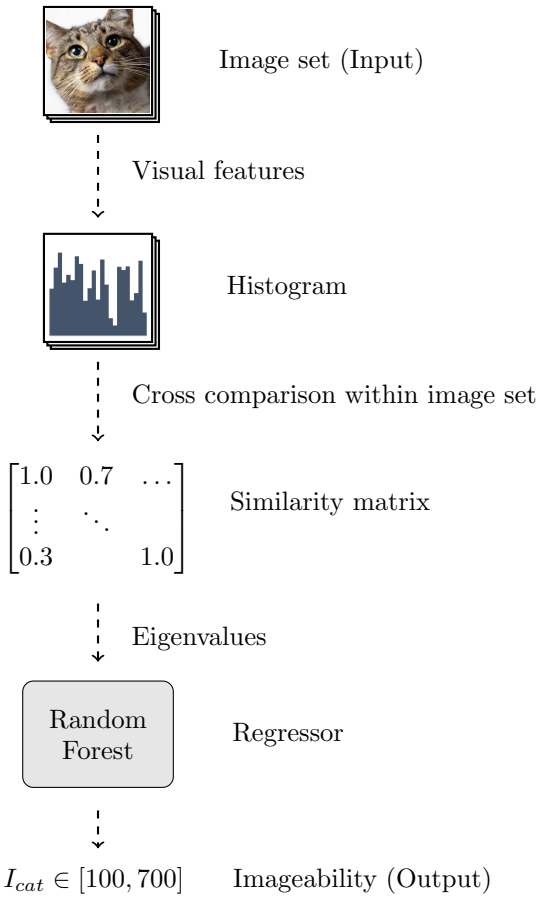

Figure 1: Flowchart of the imageability estimation. For each word, its corresponding image set gets analyzed based on its visual features. A cross comparison of these is used to create a similarity matrix between all images in the image set. Lastly, a regressor is trained to predict an imageability label.

3 Imageability estimation

In this research, we propose a direct estimation of imageability solely based on image-based data mining of Web-crawled data. The core assumption is an intrinsic relationship between imageability labels and the perceived world around us which is reflected in crowd-sourced image data from the Web. A concept, like *car* which is rather specific to us will result in a rather homogenous set of images, mostly showing very similar contents. In contrast, an abstract concept like *peace* is often much harder to visually grasp, resulting in very different and subjective mental images by different people. A crowd-sourced dataset will naturally tend to result in a much higher visual variety, showing many different contents. Thus, the core assumption is virtually equal to that of visual variety [3], yet we extend the idea to imageability by using it as a ground truth for the training.

The proposed method mines a set of images for each word, and uses the results to train a model to predict imageability. We assume a large set of images for each word, used as a basis for data mining. A variety of low-level visual features is extracted from every image. These visual features describe color distributions, edges, or gradients, within each image, and thus summarize the overall visual characteristics of the data.

To grasp the variety within the images, as discussed above, a similarity matrix between all image pairs of the same word is computed. A concrete word will tend to result in higher numbers, while an abstract word will tend to become lower numbers, due to more variety in image contents. The noisy nature of Web-based and crowd-sourced data results in many outliers, but the overall trend is expected to follow the described pattern. As the trend of similarities is the main characteristic we are interested in, the Eigenvalues of the similarity matrix are used for training. Lastly, a Random Forest regressor is trained to predict an imageability label from the corresponding set of Eigenvalues. The flow of the proposed method is shown in Figure 1.

4 Dataset creation

For this research, two types of datasets are needed.

A dictionary with imageability labels will provide ground-truth annotations for the training process. In the second step, an image set consisting of a large number of images is created for each word in the imageability dictionary.

4.1 Imageability dictionary

For the analyses, two imageability dictionaries [1][6] used for Psycholinguistics research have been extracted. Either dataset was assembled by hand through user studies. The imageability labels are represented as a value within an interval of [100,700] based on a 7-level Likert scale averaged over all test subjects, including in-between values due to the averaging process. For the experiment, the ground truth labels have been normalized to the interval of [0,100] which makes for more understandable results.

As there is no significant overlap nor contradictions in the word corpora, and both use similar methods in their creation, it seems feasible to merge both datasets and profit from the increased size.

4.2 Image sets

To create an image set for every word in the imageability dictionaries, we use the YFCC100M [8] image dataset. YFCC100M is a crowd-sourced image dataset based on the US photography social media platform Flickr[2], consisting of 100 million images up to 2014. Additional to the image URLs, the dataset provides various text-based meta annotations like a title, a description, taggings, and more, for each image.

To create a usable dataset for the evaluation, a set of images for every word in the imageability dictionary has been crawled. For each image, if a word from the dictionary is contained in either the image title, its description, or its user-tags, the image and word are regarded as related to another.

To not bias the proposed method through different similarity matrix sizes, the target is an equal number of images for every word. The first n images for every word are crawled. Due to different popularities of different concepts, many words are much harder to crawl than others. Bias through noise, misclassifications, and so on, is expected to be averaged out if n is set high enough. Notice that noise introduced through ambiguity or uncertainty is wanted in case of abstract concepts, as it is part of the core assumption.

[2]https://www.flickr.com/

5 Evaluation

First, we will discuss the actual implementation of our experiments, including the obtained datasets, and the tested visual features. Next, we will outline the results, and discuss observations and findings.

5.1 Environment

Datasets For use as an imageability dictionary, the datasets from [1][6] have been merged, resulting in a corpus of 5,108 words with imageability labels attached. As [1] only consists of nouns, there is a bias towards nouns. However, the second dataset [6] includes other parts-of-speech, spanning verbs, adjectives, adverbs, and some conjectures. For each term, YFCC100M was crawled for images, analyzing the image titles, description, and user-tags as described above. Scanning roughly the first 15% of the data set, 5000 images each for 577 words from the imageability dictionary were retrieved. The resulting dataset is split into 462 words for training and 115 words for testing.

Visual features To grasp the visual characteristics of the image sets, three visual features have been extracted from each image. First, the Color feature describes the overall color distribution of each image. They are three-dimensional histograms, with a color binning based on the HSV color space. Second, the GIST descriptor is a widely used feature for scene analyses, which encodes global gradients within the image. Lastly, SURF descriptors are extracted and used to generate a Bag-of-Words model. SURF is a local feature transformation commonly used for object detection or reconstruction. The trained Bag-of-Words model provides a histogram describing the occurrence of visually similar sub-regions in the images.

Each feature describes a different type of low-level visual characteristic. For each feature, a separate similarity matrix is computed. For the regression, the top 30 Eigenvalues of each similarity matrix were used. To round up the evaluation, a combined result, where all three sets of Eigenvalues are concatenated, will evaluate whether the visual features complement each other.

Software The prototype is implemented in Python 3.7. For visual feature extraction, OpenCV 3.2.0[3] is used, while the training uses an implementation of Random Forests from Sklearn 0.19.0[4].

[3]https://opencv.org/

[4]https://scikit-learn.org/

Table 1: Results of the experiment. While the ground truth labels are based on the Likert scale, the labels have been normalized to an interval of [0,100] to improve understandability of the results.

Feature	Correlation (1 = best)	MAE (0 = best)
Color histogram	0.56	11.78
SURF/BoW	0.55	12.53
GIST	0.45	13.09
Combined	**0.62**	**11.68**

5.2 Results

The overall results are shown in Table 1. The correlation can reach up to 0.62, while the lowest error is 11.68. The method where all visual features were combined leads to both the best overall correlation to the ground truth, and the smallest mean average error.

If converted back to the 7-level Likert scale used in the ground-truth, the error is about half a level, which means that it, in average, rounds to the *correct* bin. While each individual visual feature shows a similar correlation, the combined method correlates stronger to ground truth. This indicates, that the visual features can complement each other successfully.

6 Conclusion

In this research, we proposed a method using image-based data mining to estimate imageability labels for words. The evaluations show a mean absolute error of 11.68 and a correlation of 0.62. Therefore, we can show that the results correlate to the ground truth Lickert scale, succesfully estimating whether a term is imageable or not. The error is small enough that a mapping back to the Likert scale would round to the correct bin.

So far, most available imageability datasets are created by hand with test subjects or crowd-sourcing. This is labor-intensive and thus all datasets are small compared to the full word corpora of natural languages. Our data mining-driven method can be used to increase the vocabulary in such datasets.

In future work, we plan to look into other features, like a more high-level view on visual features to complement the currently used low-level feature set. High-level features, like actual image contents or image compositions, might give additional insights on human perception, which are not necessarily directly retrievable from lower level features. We also want to further increase the dataset, and evaluate the use of a more complex network for regression. Lastly, an evaluation across different languages would be interesting.

Acknowledgments

Parts of this research were supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research, and a joint research project with NII, Japan.

References

[1] M. J. Cortese and A. Fugett. Imageability ratings for 3,000 monosyllabic words. *Behav Res Methods Instrum Comput*, 36(3):384–387, August 2004.

[2] J. D. J. Deng, W. D. W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. 2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 2–9, 2009.

[3] M. A. Kastner, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase. Estimating the visual variety of concepts by referring to web popularity. *Multimed Tools Appl*, Published online in August 2018.

[4] G. A. Miller. WordNet: A lexical database for English. *Comm. ACM*, 38(11):39–41, November 1995.

[5] A. Paivio, J. C. Yuille, and S. A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *J Exp Psychol*, 76(1):1–25, 1968.

[6] J. Reilly and J. Kean. Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *J Cogn Sci*, 31(1):157–168, February 2010.

[7] F. Smolik and A. Kriz. The power of imageability: How the acquisition of inflected forms is facilitated in highly imageable verbs and nouns in Czech children. *J First Lang*, 35(6):446–465, October 2015.

[8] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The new data in multimedia research. *Comm. ACM*, 59(2):64–73, January 2016.

[9] M. Zhang, R. Hwa, and A. Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK*, September 2018.